798.01 / Y5

Understanding neural representations in early visual areas using convolutional neural networks

Yimeng Zhang[a], Corentin Massot[a], Tiancheng Zhi[b], George Papandreou[c], Alan Yuille[c], Tai Sing Lee[a]

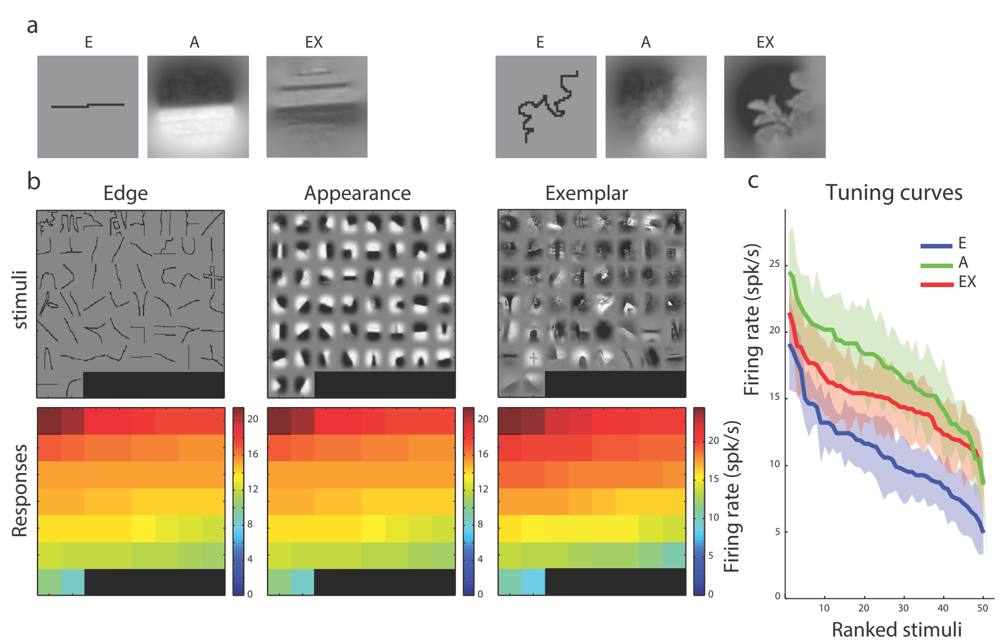
[a]Center for the Neural Basis of Cognition and Computer Science Department, Carnegie Mellon University, Pittsburgh, PA [b]Peking University, Beijing, China [c]UCLA, Los Angeles, CA

Motivations

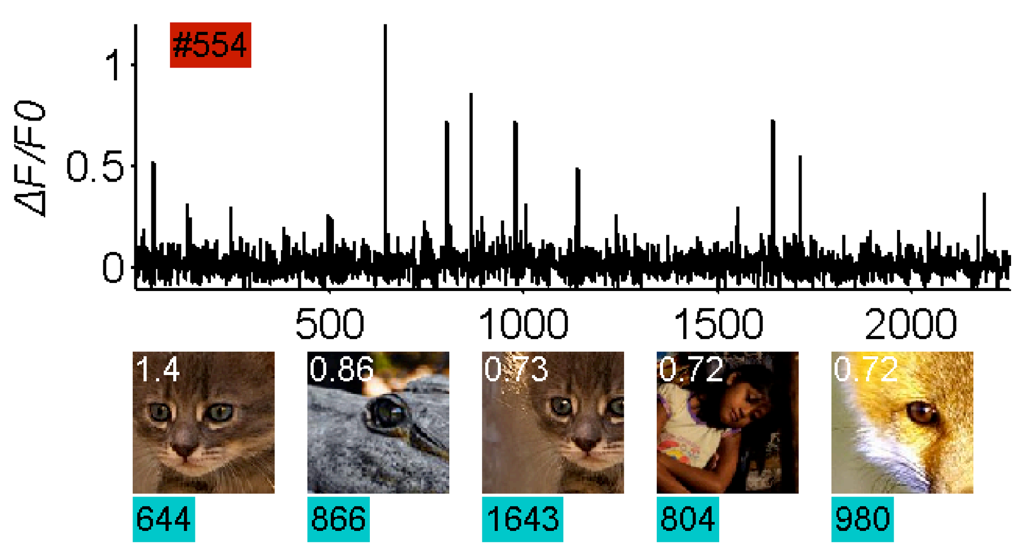
- Convolutional neural networks (CNNs) have feature representations like those in **higher** layers of the primate and human visual cortex (Agrawal et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014).
- Recent data on V1 neurons suggested that they may encode much more complex features (see poster **798.03/Y7**).
- CNN might be a useful tool to understand the encoding of complex features in **lower** layers (V1/V2) of visual cortex as well.

Images and neural data

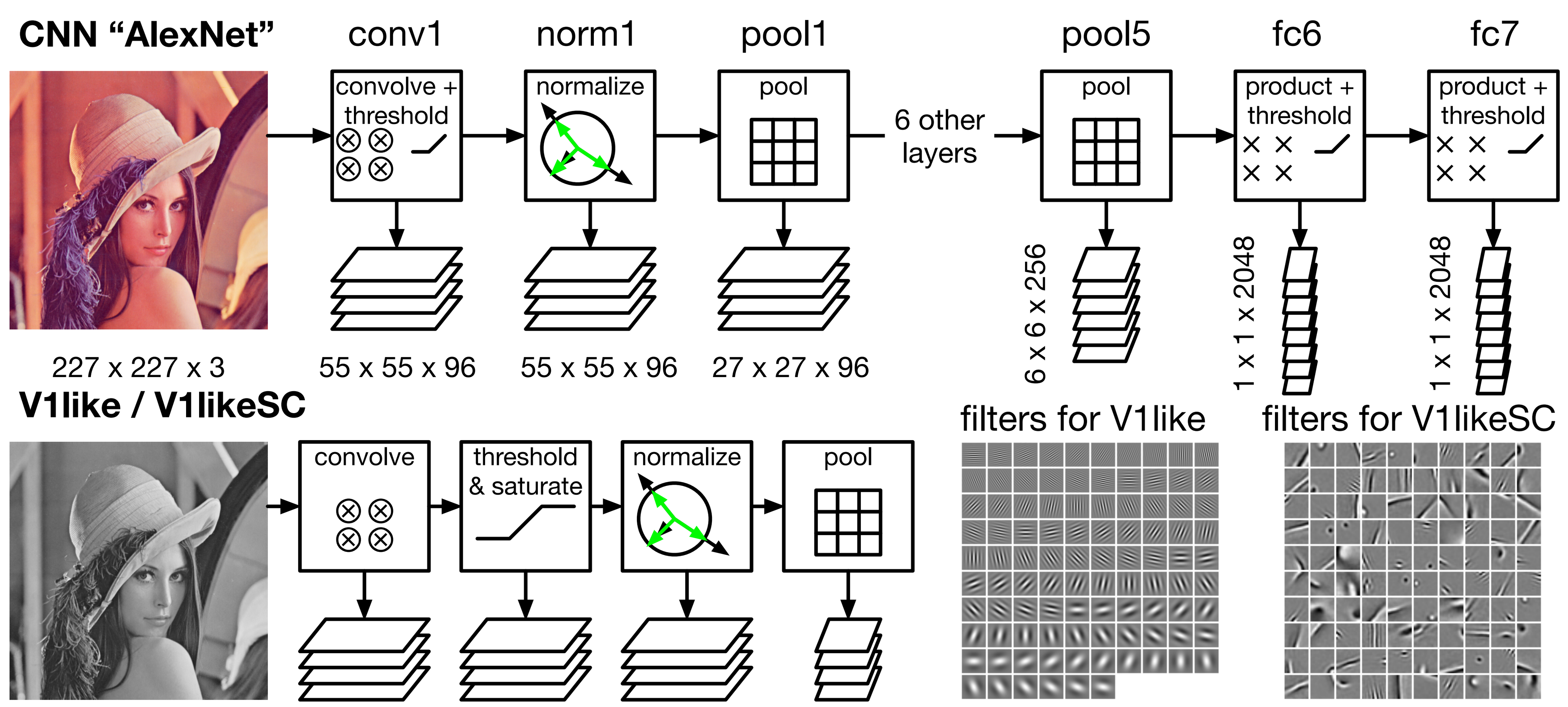
- 286 V1 and 390 V2 neurons in 2 monkeys to 150 stimuli using multi-electrode arrays.
- The 150 stimuli have 3 subsets of 50, Edge (E), Appearance (A), and Exemplar (EX), in levels of increasing complexity.



- Around 3000 V1 neurons in 3 monkeys to 2250 natural images, using calcium imaging (see poster **798.03/Y7**).

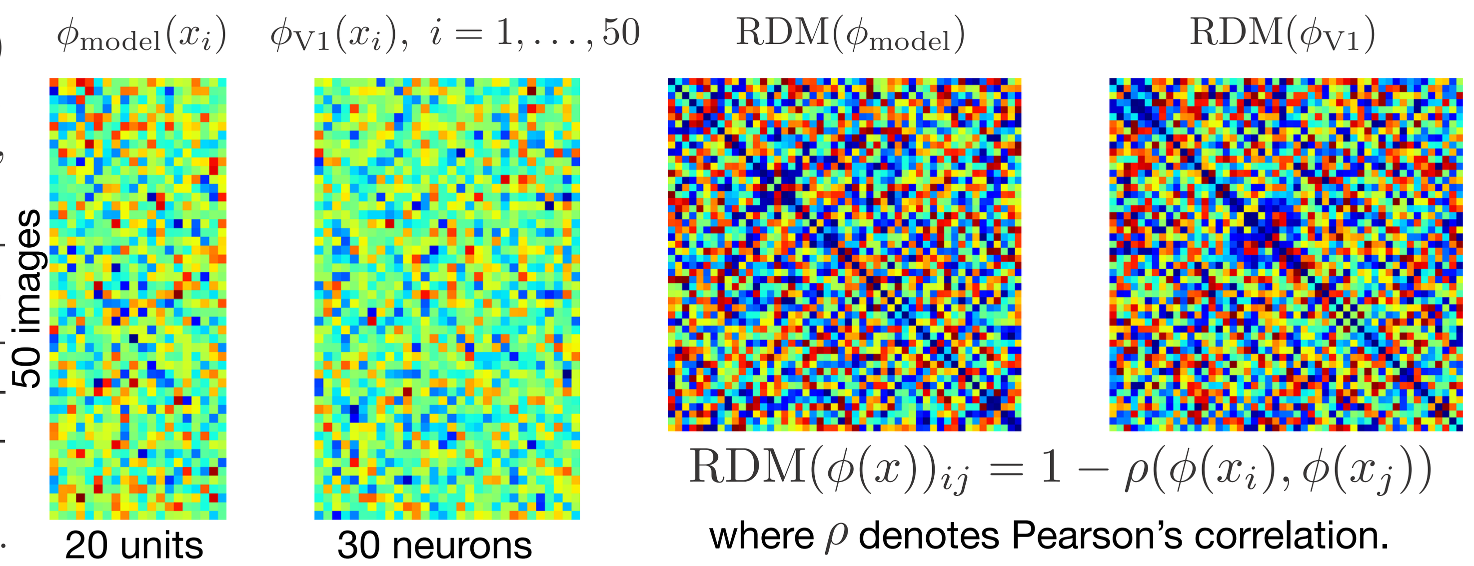


Computer vision models



Model comparison using Representational Similarity Analysis

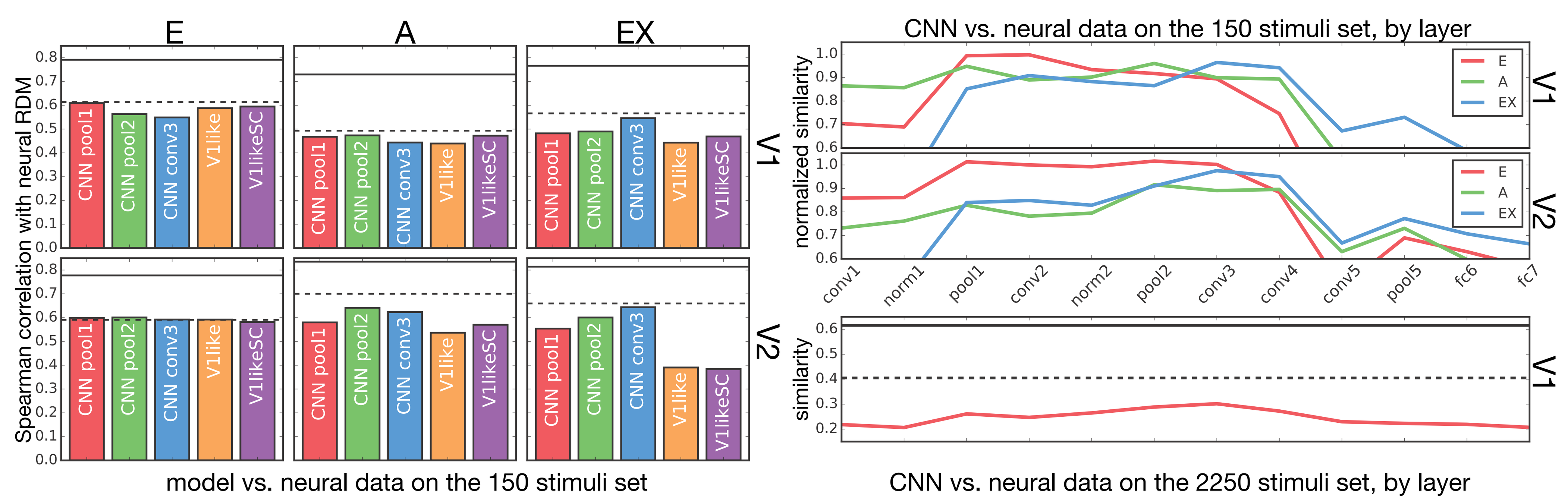
- RSA (Kriegeskorte et al., 2008) was used to compare model and neural representations ϕ_{model}, ϕ_{area} (area can be V1, V2, etc.)
- Similarity between representations is defined as the Spearman's correlation coefficient of their representational dissimilarity matrices (RDMs), each of which captures pairwise distances between images for a given representation.



$$\text{RDM}(\phi(x))_{ij} = 1 - \rho(\phi(x_i), \phi(x_j))$$

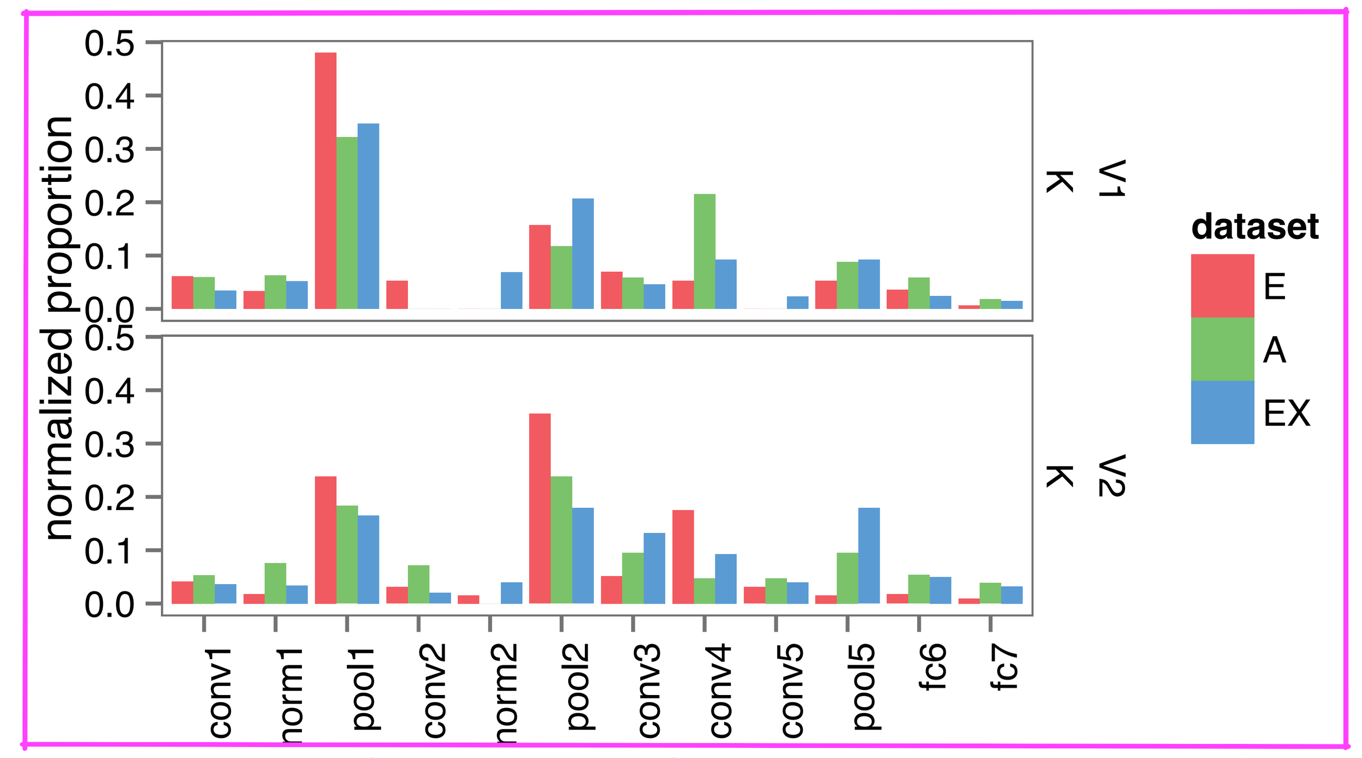
where ρ denotes Pearson's correlation.

Results



- **Left**: comparison of models on the 150 stimulus set. **Top right**: all CNN layers on the 150 stimulus set. **Bottom right**: all CNN layers on the 2250 stimulus set.
- Horizontal lines estimates the achievable similarity by computing the similarities of feature representations among different monkeys. Similar to "explainable variance".
- 150 set: CNN > V1likeSC > V1like, especially on complex stimuli (EX), and the best matching CNN layer is **stimulus dependent**, simpler stimuli (E) best explained by lower layers, and complex stimuli (EX) by higher layers.
- 2250 set: CNN is far away from achievable similarity, suggesting missing constraints in CNN.
- Higher layers in CNN can match well (even better) with V1/V2 than lower layers, suggesting complex coding in V1/V2 neurons.

Neuron matching and visualization



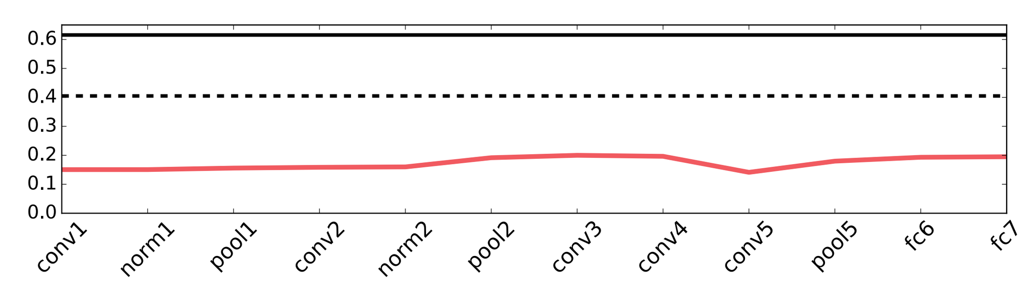
Distributions of the layers of best matching CNN units for each individual neuron.

visualization of some best matching units using deconvolution.

- Single neuron matching results were consistent with RSA: V1 matched better to pool1, V2 to pool2. Complex stimuli (EX) shifted to higher layers compared to simple stimuli (E). V2 neurons were also more correlated to higher layer CNN units than V1 neurons.
- While some neurons have visualizations consistent with the existing literature (a,b,c,e), some neurons preferred more complex features (d,f).

Why CNN performs better

- Network effects. Without normalization and pooling, V1like performed worse (not shown).
- Diverse filter shapes. V1likeSC and pool1 are better than V1like partially due to learned diverse filters compared to Gabor ones in V1like.
- Network architecture might contribute as well. On the 2250 stimulus set, higher CNN layers performed better than lower layers even with all network parameters being random.



Conclusion

- Some V1/V2 neurons may encode more complex features than previously thought.
- CNN is a good approximate model for understanding and visualizing V1 and V2 neurons.
- Future work: (1) add more biological constraints into CNN models to make CNN explain neural data better, (2) explore CNNs with heterogeneous layers, each layer with units of different complexities.

References

- Agrawal, P. et al. 2014, arXiv.org.
- Khaligh-Razavi, S.-M. & Kriegeskorte, N. 2014, PLoS computational biology, 10, e1003915.
- Kriegeskorte, N., Mur, M., & Bandettini, P. A. 2008, Frontiers in Systems Neuroscience, 2.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012, in NIPS 25, 1097–1105.
- Olshausen, B. A. 2013, in IS&T/SPIE Electronic Imaging.
- Pinto, N., Cox, D. D., & DiCarlo, J. J. 2008, PLoS Comput Biol, 4, e27.
- Yamins, D. L. K. et al. 2014, Proceedings of the National Academy of Sciences, 111, 8619.

Acknowledgments

This research is supported by IARPA MICRONS contract #D16PC00007, and NIH R01 EY022247.

Thanks to Professor Shiming Tang for the calcium imaging monkey data.

Contact: **yimengzh@cs.cmu.edu**